Topic 7: Word Embeddings

Clarissa Boyajian

2022-05-17

Download data

```
# run in console
options(timeout = 300)

# read in data (only need to run the lines below once)
download.file('https://nlp.stanford.edu/data/glove.6B.zip',
              destfile = '../data/glove.6B.zip')
unzip('../data/glove.6B.zip')
```

```
glove_data <- fread(here("data/glove.6B.300d.txt"),
                    header = FALSE)

glove_data_clean <- glove_data %>%
  column_to_rownames(var = "V1")

# convert df to matrix
glove_matrix <- as.matrix(glove_data_clean)
```

Question 1

Recreate the analyses in the last three chunks (find-synonyms, plot-synonyms, word-math) with the GloVe embeddings. How are they different from the embeddings created from the climbing accident data? Why do you think they are different?

Create synonym function

```
# take single word vector and compare to all words in all dimension
# output a similarity score
# similarity score = based on words in the 5 word windows and how they occur together
search_synonyms <- function(word_vectors, selected_vector) {

  dat <- word_vectors %*% selected_vector

  similarities <- dat %>%
    tibble(token = rownames(dat), similarity = dat[, 1])
```

```
  similarities %>%
    arrange(-similarity) %>%
    select(c(2, 3))
}
```

Calculate similary scores for "fall" and "slip"

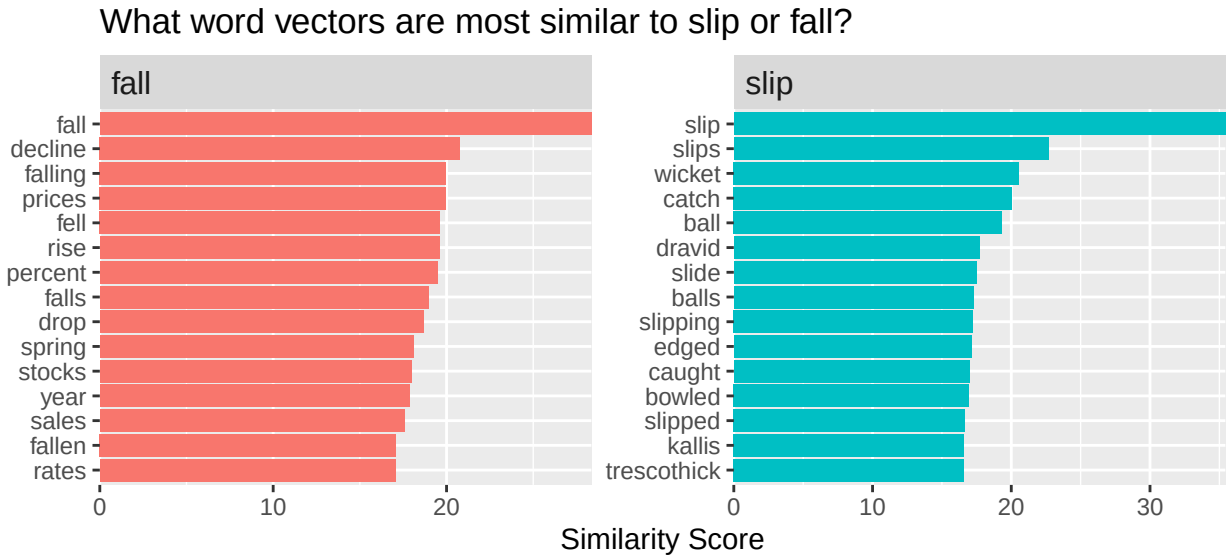
```
fall <- search_synonyms(glove_matrix, glove_matrix["fall",])

slip <- search_synonyms(glove_matrix, glove_matrix["slip",])
```

Plot Similarity Score

```
# wrangle data for plotting
plot_data <- slip %>%
  mutate(selected = "slip") %>%
  bind_rows(fall %>%
              mutate(selected = "fall")) %>%
  group_by(selected) %>%
  top_n(15, similarity) %>%
  ungroup %>%
  mutate(token = reorder(token, similarity))
```

```
# plot
ggplot(data = plot_data,
       aes(x = token,
           y = similarity,
           fill = selected)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~selected, scales = "free") +
  coord_flip() +
  theme(strip.text = element_text(hjust = 0, size = 12)) +
  scale_y_continuous(expand = c(0,0)) +
  labs(x = NULL,
       y = "Similarity Score",
       title = "What word vectors are most similar to slip or fall?")
```



The synonyms for both "fall" and "slip" are different than those created by the climbing data and have much higher similarity scores. For this data, words with high similarity scores to "fall" seem to be much more general than when we used the climbing data. There are synonyms related to falling down (such as "decline" and "drop"), there are economic terms (such as "prices" and "stocks"), and there is the opposite season ("spring"). Words with high similarity scores to "slip" appear to mostly related to cricked (such as "ball" and "wicket"). It makes sense that the synonyms and similarity scores would be different because we are no longer using climbing specific prose to create our data.

Word math

```
snow_danger <- glove_matrix["snow",] + glove_matrix["danger",]

search_synonyms(glove_matrix, snow_danger) %>%
  head() %>%
  kbl(caption = "Snow + Danger") %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 1: Snow + Danger

token	similarity
snow	57.58158
rain	40.56130
danger	40.46035
snowfall	34.84752
weather	34.37406
winds	33.96186

```
no_snow_danger <- glove_matrix["danger",] - glove_matrix["snow",]

search_synonyms(glove_matrix, no_snow_danger) %>%
  head() %>%
```

```
  kbl(caption = "Snow - Danger") %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 2: Snow - Danger

token	similarity
danger	23.31435
risks	20.22485
imminent	18.67691
dangers	17.89223
risk	17.77783
32-team	17.56241

Question 2

Run the classic word math equation, "king" - "man" = ?

```
king_minus_man <- glove_matrix["king",] - glove_matrix["man",]

search_synonyms(glove_matrix, king_minus_man) %>%
  head()  %>%
  kbl(caption = "King - Man") %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 3: King - Man

token	similarity
king	35.29707
kalākaua	26.82616
adulyadej	26.34680
bhumibol	25.87043
ehrenkrantz	25.45746
gyanendra	25.21709

Question 3

Think of three new word math equations. They can involve any words you'd like, whatever catches your interest.

```
monster_magic <- glove_matrix["monster",] - glove_matrix["magic",]

search_synonyms(glove_matrix, monster_magic) %>%
  head()  %>%
  kbl(caption = "Monster - Magic") %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 4: Monster - Magic

token	similarity
monster	24.70505
frankenstein	19.33448
raving	16.92169
700-mile	16.18631
three-headed	16.01582
monstrous	15.96151

```
red_anger <- glove_matrix["red",] + glove_matrix["anger",]

search_synonyms(glove_matrix, red_anger) %>%
  head()  %>%
  kbl(caption = "Red + Anger") %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 5: Red + Anger

token	similarity
red	51.79739
anger	47.02403
frustration	35.49240
yellow	35.31757
blue	34.93418
resentment	34.43956

```
flower_flour <- glove_matrix["flower",] + glove_matrix["flour",]

search_synonyms(glove_matrix, flower_flour) %>%
  head()  %>%
  kbl(caption = "Flower + Flour") %>%
  kable_styling(latex_options = "HOLD_position")
```

Table 6: Flower + Flour

token	similarity
flour	64.96055
flower	52.97253
vegetable	47.77346
bread	47.76887
sugar	46.18230
butter	46.10970